UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

AUTHORS GUILD, DAVID BALDACCI, MARY BLY, MICHAEL CONNELLY, SYLVIA DAY, JONATHAN FRANZEN, JOHN GRISHAM, ELIN HILDERBRAND, CHRISTINA BAKER KLINE, MAYA SHANBHAG LANG, VICTOR LAVALLE, GEORGE R.R. MARTIN, JODI PICOULT, DOUGLAS PRESTON, ROXANA ROBINSON, GEORGE SAUNDERS, SCOTT TUROW, and RACHEL VAIL, individually and on behalf of others similarly situated,

Plaintiffs,

v.

OPENAI INC., OPENAI OPCO LLC, OPENAI GP LLC, OPENAI LLC, OPENAI GLOBAL LLC, OAI CORPORATION LLC, OPENAI HOLDINGS LLC, OPENAI STARTUP FUND I LP, OPENAI STARTUP FUND GP I LLC,OPENAI STARTUP FUND MANAGEMENT LLC, and MICROSOFT CORPORATION,

Defendants.

JONATHAN ALTER, KAI BIRD, TAYLOR BRANCH, RICH COHEN, EUGENE LINDEN, DANIEL OKRENT, JULIAN SANCTON, HAMPTON SIDES, STACY SCHIFF, JAMES SHAPIRO, JIA TOLENTINO, and SIMON WINCHESTER, on behalf of themselves and all others similarly situated,

Plaintiffs,

v.

OPENAI, INC., OPENAI GP, LLC, OPENAI, LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, OPENAI HOLDINGS, LLC, and MICROSOFT CORPORATION,

Defendants.

Case No. 1:23-cv-08292-SHS;
Case No. 1:23-cv-10211-SHS

PLAINTIFFS' PARTIAL OPPOSITION TO THE OPENAI DEFENDANTS' MOTION TO SEAL

I. INTRODUCTION

Pursuant to Rule 5(B) of this Court's Individual Rules and Practices, Plaintiffs respond in partial opposition to the OpenAI Defendants' ("OpenAI's") motion for leave to file portions of the parties' discovery-related briefing, including a letter from OpenAI's Counsel to Plaintiffs' Counsel, under seal. While Plaintiffs do not oppose OpenAI's request to maintain privacy over the names of two former employees, Plaintiffs do oppose OpenAI's request to redact information about two books datasets that OpenAI previously used to "train" GPT-3 but has not used for several years.

OpenAI's compilation and use of these two important books datasets—potentially including books from notorious pirated book websites—to train its Large Language Models ("LLMS's") and decision subsequently to delete all copies of these datasets are facts to which the public should have access. OpenAI's request to redact information about these datasets, books1 and books2, should be denied for multiple reasons.

First, OpenAI has not come close to providing sufficient, specific bases in support of sealing, instead relying on the conclusory statement that material related to its LLMs, including training sources, are confidential.

Second, there is nothing commercially-sensitive about the information at issue here, given the publicly-disclosed use of books1 and books2 in training. The information about books1 and books 2 that OpenAI provided is limited. OpenAI has not disclosed the contents of these datasets, their sources, how they were created, or what particular books populate them. Instead, at issue is information about the fact that the publicly-disclosed datasets are *no longer* in use and that they were destroyed in mid-2022.

Third, information about the source material for the LLMs, especially as relates to alleged potential creation of corpuses of books derived from pirated book websites, is critical to the ultimate issues in the case, and will unavoidably be at issue over the life of this litigation.

II. RELEVANT BACKGROUND

A. Context of the Dispute

On April 12, Plaintiffs filed a letter motion seeking a discovery conference to resolve a ripe set of disputes between Plaintiffs and the OpenAI. Dkt. No. 108. Plaintiffs sought a highly relevant subset of information OpenAI submitted to the FTC pertaining to the precise issues in this case, such as the sources of training. *Id.* at 1-3. Plaintiffs also sought the names of two former OpenAI employees who created books1 and books2. Plaintiffs filed this letter motion under seal in light of OpenAI's stringent confidentiality designations.

A couple of days after Plaintiffs filed their letter motion, OpenAI provided the names of the two former OpenAI employees that Plaintiffs had been requesting for weeks, including (despite OpenAI's suggestion otherwise) via a formal meet and confer. OpenAI previously had been unwilling to provide this information, but its doing so in the wake of the briefing fully mooted this aspect of Plaintiffs' request, leaving only the FTC portion pending.

On April 16, concurrently with its opposition to Plaintiffs' letter motion, Dkt. No. 113, OpenAI filed the Motion for Leave to File under Seal and Response to Plaintiffs' Motion for Leave to File under Seal at issue here. Dkt. No. 112. OpenAI acknowledged that the FTC-related material should not be sealed, but requested that the Court seal two categories of information: (1) the identities of the two former OpenAI employees who created books1 and books2 (which Plaintiffs do not oppose), and (2) information related to the creation, use, and subsequent deletion of books1 and books2. Dkt. No. 111 at 2-3.

B. OpenAI's Declaration

To support its Motion to Seal, OpenAI filed the Declaration of Bright Kellogg, a program manager in OpenAI's legal department. There were *only two sentences* in Mr. Kellogg's short Declaration about the issues in question, one of which had a footnote that acknowledged that OpenAI had included books1 and books 2 in training data used to create GPT-3.

As Mr. Kellogg stated, "[t]he specific information OpenAI uses to train its models as well as the circumstances regarding and individuals involved in such training constitute highly confidential information. OpenAI *generall*y does not disclose publicly the sources used to train its current models [footnote] and takes steps to ensure the confidentiality of that information is maintained, particularly given the highly competitive nature of the artificial intelligence industry that spans global stakeholders." Dkt. No. 112-1 at ¶ 3 (emphasis added).

Plaintiffs do not question the competitive nature of the industry (indeed, this makes the market and/or potential market for training data important), nor do they challenge that OpenAI generally keeps training sources—including the extent to which previous models may train on current models—confidential. What Plaintiffs find wanting in these two sentences is the absence of specificity about threatened competitive harm from the fact of the destruction of books1 and books2, and any inferences (which would be only inferences) about how, whether, or to the extent the material in these datasets was contained or not in later models.

III. ARGUMENT

In deciding whether to seal or redact a judicial document, a court determines the weight of the presumption of public access for the particular documents at issue balanced against any competing interests against that presumption. *See BakeMark USA LLC v. Negron*, No. 23-CV-2360, 2024 WL 182505, at *1 (S.D.N.Y. Jan. 16, 2024). Here, the weight of the presumption of

public access is strong because the information that OpenAI seeks to seal pertains to issues at the very heart of this litigation.

Conversely, OpenAI's competing interests are weak to non-existent since OpenAI has not articulated, let alone precisely, the harm of this information being public. OpenAI's creation and use of books1 and books2 for training its LLMs is public, and OpenAI has not claimed that books1 and books2 were destroyed for a sensitive or trade secret reason, or that the destruction was somehow necessary for the functioning of the LLMs. At most, OpenAI speculates that the public may draw certain inferences about current model training, but, again, it has not shown how. It defies common sense, and in practice invites a broadening on the concept, to argue that what is *not* in the LLM models is itself a trade secret.

A. The Public is Entitled to Access to the Limited Set of Information Plaintiffs Seek to Unseal.

The weight of the presumption of public access is based on "the role of the material at issue in the exercise of Article III judicial power and the resultant value of such information to those monitoring the federal courts." *United States v. Amodeo*, 71 F.3d 1044, 1049 (2d Cir. 1995). "The presumption is at its strongest when the information at issue forms the basis of the court's adjudication" and "at its weakest in documents that have only a negligible role in the performance of Article III duties." *Coscarelli v. ESquared Hosp. LLC*, No. 18-CV-5943, 2021 WL 5507034, at *21 (S.D.N.Y. Nov. 24, 2021).

Here, the presumption of public access is strong. The information OpenAI seeks to withhold from the public pertains to the datasets used to train its Large Language Models—namely, the fact that the datasets were used, and that they have now been deleted. These facts go directly to the heart of Plaintiffs' allegations that OpenAI infringed their registered copyrights. To answer the question of whether OpenAI infringed Plaintiffs' copyrights requires the

identification of the datasets OpenAI used to train its Large Language Models. And, whether or not those datasets contain Plaintiffs' Class Works is one of the ultimate issues of this case. In fact, this particular information will likely arise countless times throughout the life of this litigation.

In addition, it is important to emphasize that Plaintiffs are not seeking the unsealing of specific information about internal OpenAI algorithms. They are seeking to unseal information about what OpenAI has already publicly acknowledged to have been fed into the algorithms. Further, third parties, including allegedly (and notably) Class Members, and not OpenAI, allegedly created the materials in books1 and books2. *See generally Grayson v. General Elec. Co.,* No 3:13-CV-1799, 2017 WL 923907 (D. Conn. Mar. 7, 2017) (drawing a distinction between an internal spreadsheet showing monies customers paid and discounts, which could be public, and pricing and profit documents, which were confidential).

The public knows that carbonated water is an input/ingredient into Coke; the public has not inferred from this the recipe for Coke itself. Plaintiffs' view is that the public—which already 'knows' that books1 and books2 (which in turn were made of materials that OpenAI did not create) were used to train at least some GPT models—can learn that books1 and books1 were destroyed. There has been no suggestion that this destruction was necessary or for a reason linked to the functioning and specifics of the LLM models themselves.

B. OpenAI's Arguments Fail to Overcome the Presumption of Public Access.

1. The procedural posture does not permit unjustified sealing.

Open AI's primary argument is that the presumption of public access is "somewhat" lower in the context of discovery disputes. However, as OpenAI acknowledges and the cases it cites make clear, there still must be specific reasons for sealing. In *Brown v. Maxwell,* 929 F.3d 41, 52 (2d Cir. 2019) (cited by OpenAI), the Second Circuit reaffirmed the "still substantial"

presumption of public access to juridical filings even about discovery. Nothing about the opinion, which in relevant part addressed privacy interests not at issue here, supports the notion that a mere assertion of competitive harm is enough. *Rand v. Travelers Indem. Co.*, No. 21-CV-10744 (VB)(VF), 2023 WL 4636614 (S.D.N.Y. July 19, 2023), OpenAI's other case, undertakes the sort of precise issue-by-issue specific analysis that OpenAI elides. *Id.* at *2-*3 (provisionally denying a request to seal a purportedly confidential email exchange about publicly-available industry guidance and about public facts, but granting a motion to seal other documents).

More broadly, the issues relating to books1 and books2 cannot be likened to a passing quarrel about a discrete document among a large universe of documents in discovery, which may or may not be relevant when the parties whittle down the issues. The use of potentially pirated books databases to create an admittedly public internal (now destroyed) dataset is a core issue in this case.

2. <u>OpenAI has not demonstrated specific harm or risk.</u>

To the extent OpenAI claims it has demonstrated specific harm via its Declaration, its argument is unsupported. "[T]he party seeking non-disclosure must make a particular and specific demonstration of fact showing that disclosure would result in an injury sufficiently serious to warrant protection; broad allegations of harm unsubstantiated by specific examples or articulated reasoning fail to satisfy the test." *Ashmore v. CGI Group, Inc.*, 138 F.Supp.3d 329, 351 (S.D.N.Y. 2015), aff'd, 923 F.3d 260 (2d Cir. 2019); *see also BakeMark USA LLC*, 2024 WL 182505, at *4.

Here, there are only two conclusory sentences, precisely the type of "broad allegations of harm unsubstantiated by specific examples or articulated reasoning" that fail to satisfy the test. *See Ashmore,* 138 F.Supp. at 351. The case on which OpenAI relies, *IBM Corporation v. Micro Focus (US), Inc.*, No. 22-cv-9910, 2024 WL 495137, at *1 (S.D.N.Y. Feb. 8, 2024), is

inapposite, as the parties *agreed* the material in question was a trade secret, and the short opinion did not address the specific materials, such as the extent of specificity of the showing of a trade secret.[1]

IV. CONCLUSION AND SUMMARY OF PROPOSED REDACTIONS

OpenAI's request to seal information about the fact of the destruction of books1 and books2, and the fact that books1 and books2 were not used (at least not used *directly*) in certain more recent GPT models should be denied.

Given OpenAI's acknowledgement that the FTC material may be unsealed, and Plaintiffs' non-opposition to the portion of the motion relating to former employee names, the import of denying the disputed portion of OpenAI's would be that the following materials would be unsealed (in addition to the briefing on the motion to seal):

First, from Plaintiffs' letter motion to compel[2]:

(a) The first and third redactions in the introductory paragraph;

(b) All redactions in the section titled "FTC Interrogatory Responses"; and

(c) Certain redactions in the section entitled "Identities of Critical Former Employees", namely, (i) "OpenAI revealed that it had destroyed all of its copies of books1 and books2," (ii) "Given that OpenAI destroyed the direct evidence of the content of books1 and books2," and (iii) "The FTC materials."

Second, Exhibit D to Plaintiffs' letter motion to compel (the letter from OpenAI's counsel to Plaintiffs that first acknowledged the destruction of books1 and books2).[3]

---

[1] *Cf. Bernstein v. Bernstein Litowitz Berger & Grossman LLP*, 814 F.3d 132, 136 (2d. Cir.2016) (affirming district court's denial of motion to seal notwithstanding parties' joint application).
[2] Dkt. No. 107 in Case No. 1:23-cv-08292; Dkt. No. 83 in Case No. 1:23-cv-10211.
[3] Dkt. No. 108 in Case No. 1:23-cv-08292; Dkt. No. 84 in Case No. 1:23-cv-10211.

Third, all redacted portions of OpenAI's April 16 opposition to Plaintiff's letter motion,[4] *except*: (a) "the identities of two former OpenAI researchers"; (b) "the former researcher's identities"; (c) "to identify the former researchers."

Fourth, all redacted portions of this Opposition brief, except those redactions made to maintain privacy over the two former OpenAI employees.

Dated: May 6, 2024

Respectfully submitted,

/s/ Rachel Geman
Rachel Geman
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
250 Hudson Street, 8th Floor
New York, NY 10013-1413
Telephone: 212.355.9500
rgeman@lchb.com

Reilly T. Stoler (*pro hac vice*)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
275 Battery Street, 29th Floor
San Francisco, CA 94111-3339
Telephone: 415.956.1000
rstoler@lchb.com

Wesley Dozier (*pro hac vice*)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
222 2nd Avenue, Suite 1640
Nashville, TN 37201
Telephone: 615.313.9000
wdozier@lchb.com

/s/ Rohit Nath
Justin A. Nelson (*pro hac vice*)
Alejandra C. Salinas (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1000 Louisiana Street, Suite 5100
Houston, TX 77002
Telephone: 713-651-9366
jnelson@susmangodfrey.com
asalinas@susmangodfrey.com

Rohit D. Nath (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1900 Avenue of the Stars, Suite 1400
Los Angeles, California 90067
Telephone: 310-789-3100

[4] Dkt No. 113 in Case No. 1:23-cv-08292; Dkt. No. 89 in Case No. 1:23-cv-10211.

rnath@susmangodfrey.com

J. Craig Smyser
SUSMAN GODFREY L.L.P.
1901 Avenue of the Americas, 32nd Floor
New York, New York 10019
Telephone: 212-336-8330
csmyser@susmangodfrey.com

/s/ Scott Sholder
Scott J. Sholder
CeCe M. Cole
COWAN DEBAETS ABRAHAMS & SHEPPARD LLP
41 Madison Avenue, 38th Floor
New York, New York 10010
Telephone: 212.974.7474
ssholder@cdas.com
ccole@cdas.com

Attorneys for Plaintiffs and the Proposed Classes

PROOF OF SERVICE VIA ECF

On May 6, 2024, I caused to be served the following document on all counsel of record via ECF.

PLAINTIFFS' PARTIAL OPPOSITION TO THE OPENAI DEFENDANTS' MOTION TO SEAL

/s/ Wesley J. Dozier